Please cite this article as:
Washington, S. and Haque, M. M. " On the commonly accepted assumptions regarding observed motor vehicle crash counts at transport system locations." In Proc. 92nd Annual Meeting of Transportation Research Board (TRB), Washington DC, USA, 2013.

On the commonly accepted assumptions regarding observed motor vehicle crash counts at transport system locations

Simon Washington
Civil Engineering and Built Environment, Science and Engineering Faculty and
Centre for Accident Research and Road Safety (CARRS-Q), Faculty of Health
Queensland University of Technology
2 George St GPO Box 2434
Brisbane QLD 4001 Australia
Tel: +61 7 3138 9990
Email: simon.washington@qut.edu.au

Md. Mazharul Haque*
Civil Engineering and Built Environment, Science and Engineering Faculty and
Centre for Accident Research and Road Safety (CARRS-Q), Faculty of Health
Queensland University of Technology
130 Victoria Park Rd
Kelvin Grove QLD 4059 Australia
Tel: +61 7 3138 4511
Email: m1.haque@qut.edu.au

*Corresponding Author

ABSTRACT
Readily accepted knowledge regarding crash causation is consistently omitted from efforts to model and subsequently understand motor vehicle crash occurrence and their contributing factors. For instance, distracted and impaired driving accounts for a significant proportion of crash occurrence, yet is rarely modeled explicitly. In addition, spatially allocated influences such as local law enforcement efforts, proximity to bars and schools, and roadside chronic distractions (advertising, pedestrians, etc.) play a role in contributing to crash occurrence and yet are routinely absent from crash models. By and large, these well-established omitted effects are simply assumed to contribute to model error, with predominant focus on modeling the engineering and operational effects of transportation facilities (e.g. AADT, number of lanes, speed limits, width of lanes, etc.)

The typical analytical approach—with a variety of statistical enhancements—has been to model crashes that occur at system locations as negative binomial (NB) distributed events that arise from a singular, underlying crash generating process. These models and their statistical kin dominate the literature; however, it is argued in this paper that these models fail to capture the underlying complexity of motor vehicle crash causes, and thus thwart deeper insights regarding crash causation and prevention.

This paper first describes hypothetical scenarios that collectively illustrate why current models mislead highway safety researchers and engineers. It is argued that current model shortcomings are significant, and will lead to poor decision-making. Exploiting our current state of knowledge of crash causation, crash counts are postulated to arise from three processes: observed network features, unobserved spatial effects, and ‘apparent’ random influences that reflect largely behavioral influences of drivers. It is argued; furthermore, that these three processes in theory can be modeled separately to gain deeper insight into crash causes, and that the model represents a more realistic depiction of reality than the state of practice NB regression. An admittedly imperfect empirical model that mixes three independent crash occurrence processes is shown to outperform the classical NB model. The questioning of current modeling assumptions and implications of the latent mixture model to current practice are the most important contributions of this paper, with an initial but rather vulnerable attempt to model the latent mixtures as a secondary contribution.

Keywords: Negative binomial regression, motor vehicle crashes, crash modelling, mixture models, unobserved spatial effects, behaviour, road safety, latent variables

1. INTRODUCTION

1.1 Background

There are numerous motivations for estimating defensible statistical models of motor vehicle crashes, including the identification of causal or contributing factors of motor vehicle crashes (with significant caveats), and the ability to identify sites performing 'worse than expected'. In current safety management practice these models are estimated from a transportation system perspective—interest centers on managing and mitigating transportation system risk in contrast to minimizing personal travel risk.

The primary use of prediction models is to estimate the expected safety performance of transportation system segments, be they highway segments, signalized intersections, roundabouts, or ramps. For example, prediction models are the basis for many fundamental principles in the Highway Safety Manual, Safety Analyst, and the Interactive Highway Safety Design Model. A major use of these models is to screen or identify potential sites for improvement, which are then audited and assessed for potential engineering or behavioral deficiencies. Often, a DOT is the agency to perform such tasks, and as such, considers engineering investments among an array of investments that may lie within the purview of other stakeholders' such as Governer's Highway Safety Office representatives.

Single equation models have dominated the literature on motor vehicle crashes, with Poisson and negative binomial (NB) regression models used in practice most often. The Poisson regression model is the basic count model for crashes [1] however, overdispersion is evident in much of the crash data where the variance of the crash frequency is greater than the mean due often to omitted variables that help to explain crash variation. Overdispersion has lead to widespread use of the Negative Binomial (NB) regression model [2]. Some applications of the NB model include investigating relationship between motor vehicle crashes and roadway geometries [3] or intersection characteristics [5], examining large truck crashes [4], exploring intersection crashes by collision types [6, 7], modeling motorcycle crashes [8], analyzing pedestrian crashes [9], investigating single and multi-vehicle crashes [10] among many others.

In the last two decades, various extensions of the Poisson or Negative Binomial model have been applied for modeling traffic crashes to further accommodate overdispersion that results from different kinds of heterogeneity [11]. For example, Chin and Quddus [12] applied a random effect NB model to treat the data in a time-series cross-section panel. Wang and Abdel-Aty [13] estimated a generalized equation model with negative binomial link to account for the correlation among the temporally and serial correlated crash data. Haque et al. [8] extended the NB and Poisson-Lognormal models to include auto-regressive correlation that account for structured heterogeneity introduced by data collection and clustering process. Mitra and Washington [14] investigated the structure of the overdispersion parameter of a NB model and reported that in the presence of small number of explanatory variables the assumption of fixed dispersion parameter is not favorable. Anastasopoulos and Mannering [15] applied a random parameter NB model to account for heterogeneity across observations by allowing some or all parameters to vary rather than being fixed. A number of studies [e.g., 16, 17] extended the NB model into zero-inflated models to take into account excess zero observations in the crash data, however, researchers [e.g., 18, 19] questioned the validity of the basic zero-state assumption in these models later. In response, Markov switching NB models were applied [20], which allow specific road entities to switch between multiple states over time. In an exploratory model fitting exercise, a finite mixture of NB models [21] that assume count data have been generated from heterogeneous populations were tested with reasonable model fit. Recently, multivariate count data models using multivariate Poisson-Lognormal regression model were applied to jointly model crash frequency at different levels

of severity [22, 23] or by collision types [24]. The multivariate Poisson-Lognormal model allows for the consideration of both overdispersion and possible correlation among different levels of data structure.

The general logic behind these single equation models—although seldom discussed—is that there is a single underlying crash occurrence process. Specifically, the assumption of underlying Poisson and NB models is that motor vehicle crashes are generated as a function of underlying known and unknown factors, where known factors include operational and geometric features of the road (e.g. AADT, lane widths, horizontal curvature, posted speed limits, etc.), and that unknown factors reveal their effects through 'extra Poisson variation'—variation not explained by the set of observed covariates. A further implication of this assumption is that the observed counts across sites, modeled say as NB, represent a single statistical distribution in derivation.

The assumption of a singular data generation process and resulting distribution is examined and questioned in this research. Knowledge we readily accept regarding crashes is consistently omitted from efforts to model and better understand crashes. Specifically, driver behavior factors such as distracted and impaired driving accounts for a major portion of crash occurrence yet are routinely omitted for lack of availability. Moreover, spatial factors—such as local police enforcement, nearby land uses such as bars and schools, and driving distractions—are also regularly omitted yet contribute to observed crash counts. By and large, these known effects are simply assumed to contribute to model error, with predominant focus on predictors related to geometric and traffic factors. How and why this presents a major problem for the use of single equation models is illustrated via hypothetical example in the following section. Then, the justification for an alternative specification to the single data generating process is described. In section 1.4 the study research objective is articulated, followed by a detailed description of a statistical model to reflect multiple sources of crashes in Section 2. The dataset used to illustrate the theoretical model of crashes is described in section 3, followed by Results in Section 4, and Discussion and Conclusions in Section 5.

1.2 Fundamental Limitations of Single Equation Models

To illustrate how a single equation negative binomial regression fails to capture sufficient detail of the crash counts at transport system locations, let's assume (with a simple yet generalizeable example) for simplicity that we seek to understand crash behavior at two urban signalized intersections (typically the interest is on many locations).The goal is to estimate the expected safety performance of these locations, then compare the observed crash counts to screen for potential problems (in practice the top x% of sites are screened for improvement). We might also wish to identify countermeasures for reducing crashes at any offending sites. To illustrate the points of this exercise (i.e. crash prediction) we pretend that we have infinite knowledge regarding the causes of crashes at the two sites.

Suppose two intersections are nearly identical in numerous measurable respects—including similar AADT, roadway geometry, posted speed limits, turning phases, cycle times, median treatments, and signage. In fact, from all measured geometric and traffic covariates they have the same expected safety performance as far as a single equation model based on them predicts. Their observed safety performance, however, has been different during the past year. Intersection A observed 9 crashes, 6 left-turn crashes and 3 rear-end crashes. The left turn crashes were largely related to a sight distance restriction, whereas the rear-end crashes were caused by drivers who were distracted via cell phones. Intersection B recorded 18 crashes, 6 of which were fatal drunk driving related crashes. Six were angle crashes caused by drivers running a red light, while the six remaining crashes were right turn on red, and related to a sight obstruction issue.

Assuming that these two intersections and their crash histories represent a plausible scenario (any number of scenarios would serve the purpose here), consider the following important, critical, and constructive insights:

1. Any expected safety performance function (SPF) estimated using the crash counts of these two intersections (forgetting sample size related issues for the moment) will use the crash counts 9 and 18 respectively, and attempt to statistically relate these to the observed geometric and traffic covariates.
2. The typically observed covariates, being the same for both sites, will result in a safety performance function that predicts approximately 13.5 crashes per year for this type of intersection.
3. The expected safety performance of 13.5 is generally interpreted to mean that crashes above this amount conditional on the covariates (ignoring for the moment potential site selection bias) are indicative of potentially correctable sites (with larger differences being more 'correctable').
4. Intersection B would be identified as a potential high risk site, as it recorded 18 crashes but should be recording, on average, 13.5.
5. At intersection B in reality (recall we are omniscient in this example), 6 of the crashes may be correctable from engineering improvements and 6 through enforcement (perhaps automated). The remaining 6 impaired driver crashes will likely require legal system changes or enforcement activities to correct.
6. In reality only 6 of the crashes at intersection B are preventable through engineering investments—the same as the number at intersection A (the 6 left turn crashes).
7. The estimated SPF is incorrect for both intersections—because the crashes caused by behavioral problems are not in fact a function of intersection covariates at these intersections. Considering only geometric and traffic features, each intersection should produce about 6 crashes per year, with additional crashes caused by predominately non-engineering factors.
8. Despite our best of efforts, the SPF did not help to identify the true intersections in need, failed to identify the expected safety performance, and co-mingled crashes caused by a variety of factors. The problem stems from the fact that crashes are caused by a variety of causes, not just operational factors.

Of course any number of hypothetical scenarios could be constructed to illustrate this point. The realistic scenario, though hypothetical, does raise some very interesting and important points regarding the collective approach that has to date been widely adopted to model crashes at intersections, the negative binomial regression (and variations) assuming a single crash occurrence process. These concerns include:

1. Crashes that result from behavioral factors—that are predominately unrelated to geometric factors—such as driver distractions, impaired driving, and the like, are often not correctable through engineering improvements alone, but regularly contribute to observed crash counts. Most reputable studies [e.g., 25] suggest that more than 50% of crashes are primarily the result of human error (and only weakly related to geometric and traffic factors), and many suggest that human factors are about 90%.
2. Crashes related to unobserved spatial factors, such as proximity of bars or schools, chronic glare conditions, heterogeneous driver population effects (e.g. predominance of young or older drivers), and others are generally not correctable through engineering improvements alone either. Many studies [e.g., 26, 27] have postulated the existence of such effects and have shown evidence for them; however, their overall contribution is generally not known with certainty.

3. The inclusion of crashes that are not generally correctable through engineering improvements in observed crash counts tends to result in SPFs that over-estimate the number of crashes we should expect to see at a site due to geometric factors and traffic.
4. Because the factors that capture behavioral and spatial effects are omitted from SPFs (they are generally unavailable), their effects are mistakenly attributed to included correlated variables, resulting in biased parameters.
5. Since the contribution of behavioral and spatial effects across sites is somewhat random (more on this later), their presence severely hinders the ability to screen sites accurately.

1.3 A more realistic alternative to the single crash generating process

It was described and illustrated via example previously that observed crashes are not generated by a single underlying process, but instead arise as the result of three separate processes that lend themselves to greater insight and understanding of crash causation. These three separate processes are crashes that arise from behavioral factors, from geometric and operational factors, and from unobserved spatial effects. Moreover, the three separate crash generating processes lend themselves—with the help of some exogenous information—to statistical modeling, although requiring greater complexity than single equation approaches.

It is postulated here that three processes give rise to three separate crash counts and associated probability distributions, which when summed at a site, constitute the total observed crash count on transportation networks. Specifically, site geometric and operational factors will produce crashes that correspond to an observed distribution across sites, as do behavioral factors (e.g. driver distraction, fatigue, etc.), and spatial factors (e.g. the effect of a drinking establishment on impaired driving crashes locally or the effect of an elementary school on pedestrian involved crashes locally). More detail on each of these distributions is provided below.

Geometric and Traffic Influences on Crashes

The discussion begins with this category—the set of observed crash frequency influences—because these factors dominate current models and thus are familiar. Much research has established that measureable geometric and operational features of transportation system segments influence the frequency of crashes observed at a location. At intersections the phasing, channelization, and median treatments tend to have significant effects. On segments the shoulder width and treatment, lane widths, and surface condition all are known to influence safety. Of course, the exposure of the driving population to such features is always a dominate factor influencing crash frequencies, either measured by AADT (segments) or by entering traffic volumes (intersections).

Spatial Influences on Crashes

This category tends to arise from a variety of influential factors. Many spatial factors known to affect safety are not typically measured or observed. Examples include impaired drivers leaving a local drinking establishment, local pavement conditions, local distractions (billboards, glare, signs, etc.), proximity to a university (pedestrian and bicycle traffic), or collisions with animals (that are in abundance in proximity to certain rural locations). These factors tend to influence safety in systematic ways associated with specific locations, and contribute to crash counts observed on the system. Of course exposure is again a dominant factor to help estimate these effects of these influences, and location is important.

Behavioral Influences on Crashes

Random effects—as observed by the transport network—are simply a function of exposure. Most random effects are related to human factors issues that are unobserved or unknown at the time of the crash. Examples include in-vehicle distractions (cell phone, changing radio, eating food, day-dreaming, fatigue, talking with passenger, etc.) and rare random events such as striking debris in the road, impacting domestic animals, mechanical breakdowns, etc. From a system point of view these events are largely unpredictable given system characteristics—either geometric or operational—and thus do not exhibit patterns as a function of system characteristics.

Figure 1 shows a bar diagram for the crash components described previously, abbreviated "Observed", "Unobserved", and "Random". The figure depicts a hypothetical set of 30 sites with observed total crash frequencies on the vertical axis. In theory, random crashes—shown in black—contribute to the total crash sum, as do the other components. Some sites, for example site 8, may only have random crashes occurring, while sites 18 and 20 only have crashes that result from unobserved spatial effects. Some sites may have all three components, such as site 16. The critical point here is that theoretically these three categories of crashes contribute to the total observed at a site.

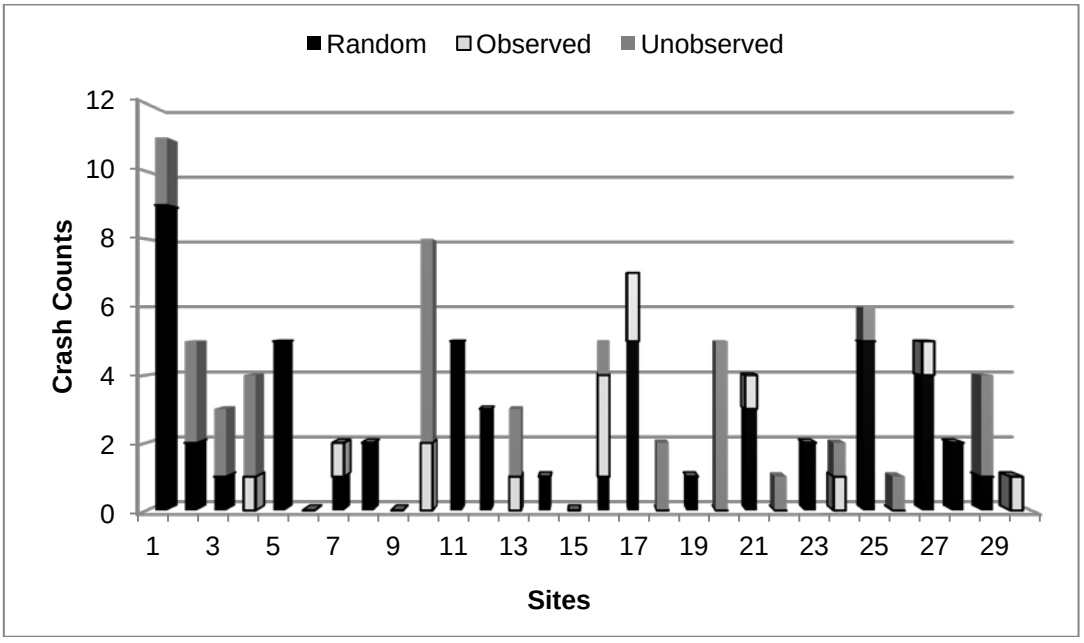

FIGURE 1 Hypothetical sum of crash counts from observed, unobserved and random effects

Another important point is illustrated in Figure 2. This figure shows components and the totals (sum of all 3 components). Past research has focused predominately on modeling the total crashes across sites (or all rear-ends, fatals, etc.) shown in hatched, whereas the ability to isolate the individual components might provide a distinct comparative advantage. For example, site 1 appears to be the worst site in the sample based on total counts, yet the number of crashes due to observed effects is zero and the random effect is large (implying perhaps human factors issues). Thus, it would not be useful to examine site 1 for geometric or operational deficiencies. Site 20, in contrast, consists of crashes that result from unobserved spatial effects. This hypothetical set of sites serves to highlight the problem that evades the profession currently: we assess underlying safety and develop performance expectations

based on contributions of various elements that when combined challenge our ability to effectively identify contributing factors.

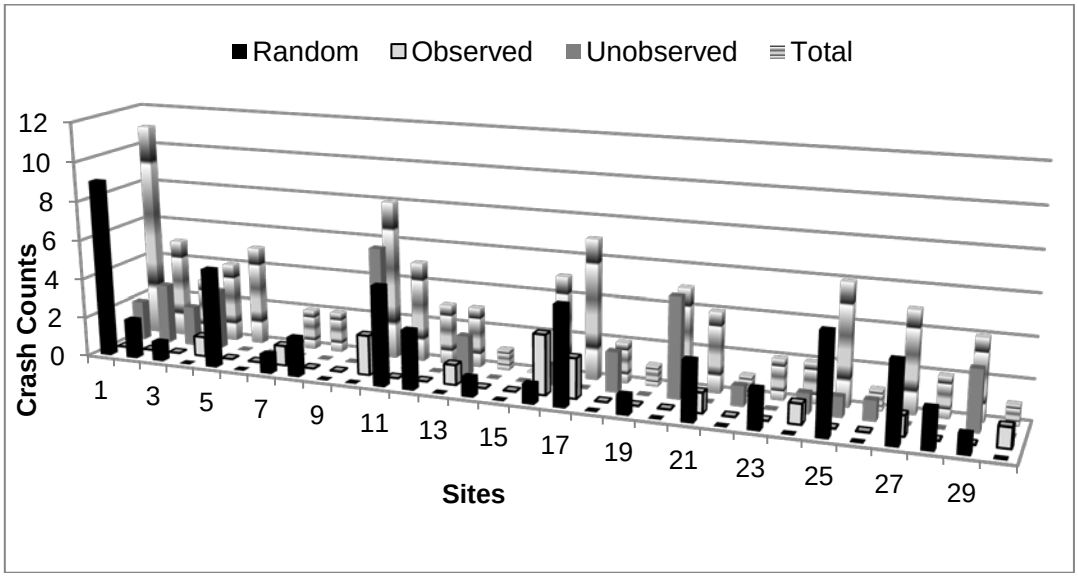

FIGURE 2 Comparison of hypothetical observed and total effects across sites

1.4 Research Objective

This study formulates and presents a methodology that accounts for the mixture of the three crash occurrence processes described previously to develop a rigorous, more accurate safety performance function for a site. To model the three crash components, separate univariate Poisson or Negative Binomial models are developed and combined. Prediction performances of the state of the practice single equation negative binomial model and the proposed three components mixture model are compared.

2. METHODOLOGY AND ANALYTICAL APPROACH

This section presents the formulation of the proposed mixture model that accommodates three crash components. Specifications for models are briefly discussed, followed by a description on the model estimation technique used for the three component mixture model. Finally, two goodness-of-fit criteria used to compare the performance of the models are described.

2.1 Model Development

Count data modeling techniques are commonly used for crash frequency analysis since crash occurrence often follows a Poisson process. Derivatives of the Poisson process are the Poisson regression model and Negative Binomial (NB) regression model [28]. The NB model, developed to overcome the 'mean equal to variance' assumption of the Poisson regression model, dominates the literature.

2.1.1 Single Equation Negative Binomial Model

Let's assume that Y_i represent the observed crash frequency for intersection *i*, and Y_i follows a Poisson process with the Poisson mean μ_i:

$$\mathbf{Y}_i \sim Poisson(\mu_i)$$

In the Poisson regression model, the mean of the Poisson is structured as follows:

$$\log(\mu_i) = \boldsymbol{\beta}\mathbf{X}_i \tag{1}$$

where $\boldsymbol{X}_i$ is vector of covariates representing site-specific attributes, $\boldsymbol{\beta}$ is a vector of unknown regression parameters. To accommodate the overdispersion, researchers have proposed the inclusion of gamma-distributed error term in the parent Poisson model that formulates the Negative Binomial regression model as follows:

$$\begin{aligned} \mathbf{Y}_i &\sim NB(\mu_i, \varphi) \\ \log(\mu_i) &= \boldsymbol{\beta}\mathbf{X}_i + \varepsilon_i \end{aligned} \tag{2}$$

where ε_i is the model error independent of all covariates. It is assumed that $\exp(\varepsilon_i)$ is gamma distributed ($Gamma \sim (\nu, \nu)$) with mean 1 and variance $1/\nu$ for all *i*. Hence, the dispersion parameter, $\varphi = (1/\nu)$ of the NB model accommodates extra variation in the crash data. When φ is equal to zero, the NB model reduces to a Poisson regression model.

The above formulation of the NB model is simplistic and requires sophistication when capturing unique features of crash data. Numerous studies [e.g., 7, 29] have adopted a logarithmic transformation of the exposure variables, i.e. major and minor road traffic flows. This transformation allows a nonlinear relationship between traffic flows and crashes and constrains predictions such that there are no expected crashes in the absence of exposure. The model specification with logarithmic exposure variable is:

$$\mu_i = \alpha_0 F_{1i}^{\alpha_1} F_{2i}^{\alpha_2} \exp \sum \beta_j X_{ij} \tag{3}$$

where F_{1i} and F_{2i} are respectively major and minor road flows for intersection *i*; X_{ij} are variables describing road geometry and traffic information; α_0, α_1, α_2, and β_j are estimated regression parameters.

2.1.2 Mixture Model with 3 Crash Occurrence Processes

As described previously, let's consider Y_i arises from three separate crash occurrence processes. They are: 1) observed network and operational features, 2) unobserved spatial factors, and 3) behavioral or random influences. Then, Y_i can be assumed to include three separate density functions such that

$$\mathbf{Y}_i \sim NB(\lambda_i = \sum_{k=1}^{3} \mu_{ik}, \phi_k) \tag{4}$$

where is λ_i the crash mean for i^{th} entitiy, μ_{ik} is the crash mean for i^{th} entity generated from k^{th} crash occurrence process, and ϕ_k is overdispersion parameter of the k^{th} crash occurrence process. Let's assume $\boldsymbol{\theta}$ = $(\theta_1, \theta_2, \theta_3)'$ is the mixing proportion whose elements sum to unity.

$$\begin{aligned}\mu_{i1} &= \theta_1 \lambda_i \\ \mu_{i2} &= \theta_2 \lambda_i \\ \mu_{i3} &= \theta_3 \lambda_i \\ \sum_{k=1}^{3} \theta_k &= 1\end{aligned} \tag{5}$$

The model specification for the observed network and operational features is the same as single equation NB model.

$$\begin{aligned}\mathbf{Y}_{i1} &\sim NB(\mu_{i1}, \phi_1) \\ \mu_{i1} &= \alpha_0 F_{1i}^{\alpha_1} F_{2i}^{\alpha_2} \exp \sum \beta_j X_{ij}\end{aligned} \tag{6}$$

where ϕ_1 is the overdispersion parameter for the crash occurrences related to observed network and operational features, X_{ij} is the j^{th} road geometry and traffic related variable for i^{th} intersection, and other parameters are as previously defined. The model specification for unobserved spatial factors is:

$$\begin{aligned}\mathbf{Y}_{i2} &\sim NB(\mu_{i2}, \phi_2) \\ \mu_{i2} &= \alpha_0 F_{1i}^{\alpha_1} F_{2i}^{\alpha_2} \exp \sum \beta_j W_{ij}\end{aligned} \tag{7}$$

where ϕ_2 is the overdispersion parameter for the crash component related to unobserved spatial factors, W_{ij} are variables related to spatial attributes for i^{th} intersection. The specification for the third crash occurrence process is:

$$\begin{aligned}\mathbf{Y}_{i3} &\sim NB(\mu_{i3}, \phi_3) \\ \mu_{i3} &= \alpha_0 F_{1i}^{\alpha_1} F_{2i}^{\alpha_2}\end{aligned} \tag{8}$$

where ϕ_3 is the overdispersion parameter for the crash component related to random influences. The mean structure for the random influences is simply a function of the exposure variables, i.e., major and minor road traffic flows.

2.2 Modeling Methodology

In practice it is unusual to know the proportion or frequency of crash types that contribute to an observed overall crash frequency at a site. In other words, the mixing proportions, θ_k's of the proposed 3-component mixture model are unknown, and cannot be determined without additional exogenous information. To circumvent this problem for the time being, a simulation-based estimation technique is applied to account for the three crash occurrence processes. The relative weights of the three crash occurrence components have been extracted from the literature. For example, a number of reliable studies [e.g., 25, 30] have reported that human factors are the main or primary contributor to crashes. Human factors like recognition errors, decision errors, critical non-performance, and action errors solely represent more than half of the crashes, whereas roadway, traffic and environmental factors in combination with driver and vehicle factors represent about one-third of crashes. These findings serve as a guideline to assign a distribution of weights as follows:

$$\begin{aligned} &\text{Observed network features:} && \theta_1 \sim U[0.3,0.4] \\ &\text{Unobserved spatial factors:} && \theta_2 \sim U[0.1,0.2] \\ &\text{Random or behavioral influences:} && \theta_3 = 1-(\theta_1+\theta_2) \end{aligned} \tag{9}$$

Equation 9 reflects an analysis where overall crashes are randomly and uniformly drawn from the observed network features between 30-40%, from unobserved spatial factors between 10-20%, and random or behavioural influences the remainder. Using these distribution weights, observed crash counts are assigned to three frequency counts for each site *i*. To assess the consistency and sensitivity of results to assumed weights, five different distribution proportions and corresponding models are estimated.

2.3 Goodness-of-fit

As always models are compared on some global criterion in addition to comparison of parameter estimates and standard errors. In this study two common prediction-based model selection criteria applied are: 1) mean squared predictive error (MSPE), and 2) predictive loss criteria (PLC). Suppose, ξ_i and ς_i are mean and variance of maximum likelihood estimation (MLE) based crash prediction for site *i* using asymptotic normality based on large samples. Then the MSPE is calculated as follows:

$$MSPE = \sum_{k=1}^{3}\left[\sum_{i=1}^{N}(Y_i-\xi_i)^2 \Big/ N\right] \tag{10}$$

where *k* denotes the crash occurrence processes, Y_i be the observed data and *N* is the number of observed sites. The MSPE relies on the mean of predictions and does not take into account the variance of predictions. In contrast, the PLC [31] also includes the variance of predictions and hence might be a more informative model section criterion, where

$$PLC = \sum_{k=1}^{3}\left[\sum_{i=1}^{N}\varsigma_i + [w/(w+1)]\sum_{i=1}^{N}(Y_i-\xi_i)^2\right] \tag{11}$$

and w is the weight factor. A large value of w puts more weight on the match between predicted and observed data. By assuming an infinite value for w, as used by Haque et al. [8], equal weights have been put for variance and mean differences to calculate the PLCs in this study. Models with relatively lower MSPE and PLC values are regarded as superior models in terms of statistical fit. It is recognized, however, that statistical fit should never be the sole criterion for preferring one model to another.

3. DATASET FOR ANALYSIS

To test the feasibility of this proposed model, crash data from rural intersections in 38 counties in the state of Georgia from 1996 and 1997 were used. This dataset had been extensively examined in past research [7, 14, 24] and consisted of 165 rural intersections on two-lane roads, including 51 signalized and 114 unsignalized intersections. An intersection crash was defined as any crash occurring at the intersection or within 76m (250ft) of the intersection along the major and minor roads. Using this definition a total of 837 crashes were recorded, including 345 at unsignalized and 492 at signalized intersections.

A series of explanatory variables describing roadway characteristics, intersection characteristics and geometry, traffic volumes for major and minor roadways were extracted from road characteristics files, aerial photographs, and geographic information system (GIS) roadmaps. Furthermore, Digital Orthophotography Quarter-Quadrangles (DOQQs) aerial photos from 1994 and 2000 were overlapped with GIS roadmaps to extract information regarding intersection angle and degree of horizontal curvature [7]. Major and minor roadway related variables include traffic flows in AADT, median width, shoulder width, provision of right-turn and left-turn lane, roadside hazard rating, number of driveways, lighting condition, road terrain condition, speed limit, and sight distance. Intersection-related variables include provision signalization condition, ratio of grade changes in major and minor roadways, and intersection angle. A detail description of variables along with descriptive statistics can be found in Kim et al. [7]. In addition to these, a spatial variable of county level population density was derived from the US census Bureau website for the 2000 calendar year and converted into an indicator variable, where 1 represents a population density greater than the mean (286 person/square km), and 0 represents population densities less than this. All the explanatory variables are centered and standardized before input to the proposed statistical models.

4. RESULTS

This section compares the performance of the alterative models. All the models are estimated employing the classical maximum likelihood estimation (MLE) technique. The best-fit model is selected using standard goodness-of-fit criteria and logical defensibility.

Models are estimated for a single equation NB model (the current 'state of the practice' for modeling crashes) and mixture of 3-component separate regression models described previously. The mixture model is estimated across five different randomly drawn weight distributions. Goodness-of-fit statistics for these models are presented in Table 1. On global goodness of fit measures, the mean squared predictive error (MSPE) for the single equation NB model is 23.2, while the MSPE for the 3-component mixture model is about 10. The predictive loss criterion (PLC) for the NB model is 3832.7 and about half of this amount for the 3-component mixture model. Clearly, the mixture model significantly outperforms the single equation model and supports, at least statistically, that a process whereby three separate underlying crash occurrence processes represent observed crashes is an entirely plausible, well fitting model. The goodness-of-fit statistics of the five simulation runs of the 3-component mixture model are quite similar. Among the five simulation runs, the 3rd trial shows a slightly better fit than the others with MSPE of 10.11 and PLC of 1682.2. Not shown in the table, the MSPE for observed network features, spatial factors, and random influence components were 2.9, 0.8, and 6.4 respectively. And the corresponding estimates of PLCs across the three components were 482.6, 140.4, and 1059.2 respectively.

While an omnibus comparison of models supports the three component model, a careful assessment of the theoretical appeal and justification is needed.

TABLE 1 Comparison between Single Crash Occurrence Process Model and 3-Components Mixture Model (Estimations by MLE)

Model Selection Criteria	NB Model	3-Components Mixture Model				
		1st trial*	2nd trial	3rd trial	4th trial	5th trial
Mean Squared Predictive Error (MSPE)	23.20	10.30	10.33	10.11	10.43	10.51
Predictive Loss Criteria (PLC)	3832.68	1713.00	1718.72	1682.16	1734.11	1748.84

*each trial represents a random draw of weight distributions

4.1 Single Equation NB Model

Estimation results of the single equation NB model by the MLE technique are presented in Table 2. The best-fit model consists of six explanatory variables with a *pseudo-R^2* of about 0.12. The overdispersion parameter and explanatory variables are statistically significant at the 5% significance level. Significant variables include traffic flow on the major and minor roads, median width, provision of a right-turn lane, number of driveways, and lighting condition of the major road. All the variables have plausible signs and magnitude. For example, traffic volumes increase crash risk, whereas the presence of lighting and increasing median width improve safety. The magnitudes of parameter estimate in Table 2 are different from those provided in Kim et al. [7], because the NB model here is estimated with standardized covariates. None the less, the effects of model variables on safety are similar.

TABLE 2 Single Equation Negative Binomial Regression Model

Variables	Estimate	SE	z-statistic	p-value
Constant	1.315	0.067	19.48	0.000
Log of AADT on the major road	0.408	0.080	5.12	0.000
Log of AADT on the minor road	0.317	0.077	4.10	0.000
Major road median width	-0.224	0.070	-3.21	0.001
Right-turn lane indicator in the major road	0.262	0.069	3.80	0.000
Number of driveways on the major road	0.237	0.080	2.95	0.003
Lighting indicator on the major road	-0.192	0.078	-2.47	0.014
Dispersion parameter, φ	0.414	0.079		0.000
No. of Observations	165			
Log-likelihood at convergence	-394.521			
Pseudo-R^2	0.119			

4.2 Mixture of 3-Component Separate Regression Models (MLE)

Table 3 through 5 present the estimation results from the separate regression models for the three underlying crash components: observed network effects, spatial effects and random or behavioral effects. Estimation results of the best-fitting 3rd simulation trial are reported here (see Table 1).

TABLE 3 Negative Binomial Regression Model Results: Observed Network Effects

Variables	Estimate	SE	z-statistic	p-value
Constant	0.256	0.080	3.18	0.001
Log of AADT on the major road	0.471	0.094	5.03	0.000
Log of AADT on the minor road	0.305	0.074	4.14	0.000
Major road median width	-0.146	0.063	-2.32	0.020
Right-turn lane indicator in the major road	0.246	0.060	4.10	0.000
Number of driveways on the major road	0.210	0.077	2.72	0.006
Lighting indicator on the major road	-0.241	0.076	-3.16	0.002
Dispersion parameter, ϕ_1	0.103	0.058		0.010
No. of Observations	165			
Log-likelihood at convergence	-243.438			
Pseudo-R^2	0.186			

The NB model for observed network features component retains the same six explanatory variables as for the Single NB model and yields a *pseudo-R^2* of 0.19. As

hypothesized, this model component captures the safety effects of operational and geometric features. Similar to prior studies, AADT on major and minor roads are positively associated with intersection crashes implying that increased traffic volume or exposure to risk are associated with increased crashes. Median width on major roads is negatively associated with crashes as physical separation of travel directions is likely to improve safety [e.g., 12, 32]. The provision of right-turn lane on the major road is associated with an increase in the number of crashes. An exclusive right-turn lane is generally installed at intersections where the volume of right-turn traffic or the number of right-turn crashes is high, and thus this effect may represent an endogeneity problem as described in Kim et al. [7]. An alternative explanation is that a right-turn lane represents a significantly larger proportion of turns at the intersection relative to other sites without right turn lanes, and thus more potential angle conflicts and crashes are possible. The number of commercial driveways along the major roads is positively associated with intersection crashes. Access points close to the intersection are likely to increase the complexity of traffic movements and generate an increase in conflict opportunities. The lighting condition is negatively associated with intersection crashes, suggesting that greater visibility at night decreases the intersection crashes and hence increase safety relative to similar sites without lighting.

TABLE 4 Poisson Regression Model Results: Unobserved Spatial Component

Variables	Estimate	SE	z-statistic	p-value
Constant	-0.538	0.113	-4.76	0.000
Log of AADT on the major road	0.482	0.122	3.97	0.000
Log of AADT on the minor road	0.364	0.086	4.25	0.000
Population Density	0.172	0.086	1.99	0.046
No. of Observations	165			
Log-likelihood at convergence	-162.125			
Pseudo-R^2	0.177			

Model estimates for unobserved spatial effects are presented in Table 4. For spatial effects, a Poisson regression model is fitted since the overdispersion parameter for a NB formulation was not significant. As shown in Table 4, the Poisson regression model yields a *pseudo*-R^2 of 0.18 and retains three explanatory variables: AADT of major road, AADT of minor road, and county-level population density. Population density is a statistically significant predictor of unobserved spatial effects. Population density serves yields increased explanatory power of this model, and reflects a greater likelihood of unobserved spatial effects where population density is greater, not surprisingly. In a spatial analysis in southeastern Michigan [33], population density was also a significant predictor for motor vehicle crashes. Recall that this model component is meant to capture the effects of contributors to crashes that are related to fixed points in space but yet are not readily measured for use in crash modeling—and thus are unobserved. As one would expect, for unobserved spatial influences to contribute to crash occurrence will require vehicles on the road and nearby space related attributes—both of which are reflected in this model.

TABLE 5 Negative Binomial Regression Model Results: Random Influence Crash Component

Variables	Estimate	SE	z-statistic	p-value
Constant	0.664	0.077	8.62	0.000
Log of AADT on the major road	0.523	0.086	6.07	0.000
Log of AADT on the minor road	0.312	0.074	4.23	0.000
Dispersion parameter, ϕ_3	0.335	0.083		0.000
No. of Observations	165			
Log-likelihood at convergence	-304.838			
Pseudo-R^2	0.117			

The NB model estimates for the random or behavioral influences on crash occurrence are shown in Table 5. Recall that this model component is meant to capture the effects of behavioral issues (e.g. distraction, fatigue, impaired driving, inattention, etc.), the majority of which are rarely if ever measured and included in crash models—and thus appear 'random' from the point of view of the transportation system. The random influence model is specified as a function of exposure only, with the logical hypothesis that apparently random contributors to crashes will increase with increasing exposure. The NB model estimates a *pseudo-*R^2 of 0.12 with a significant overdispersion parameter of 0.34 and explanatory variables significant at the 5% level.

5. DISCUSSION AND CONCLUSIONS

This study argues from a theoretical standpoint that observed crashes are not generated by a single crash occurrence process and instead arise as a result of three separate processes including observed network influences, unobserved spatial influences, and behavioral influences that appear random from the system point of view. This argument is based on decades of evidence articulating the contribution of these three distinct sources of crashes, and the logical progression that when we observe crash counts on network sites we are observing some unknown sum of these separate components.

Based on this more detailed view on how crashes accrue, we formulate a statistical model that takes into account these processes. The latent mixture model—a model that represents a sum of an unobserved mixture of distributions—is formulated statistically, stating the underlying assumptions. Through analysis of rural intersections in Georgia, the mixture model of the three crash components model is estimated using the maximum likelihood estimation technique. The goodness-of-fit statistics of the mixture model are compared with single equation negative binomial model.

Model estimates reveal that the proposed theoretically motivated mixture model of three crash components provides a vastly superior statistical fit compared to the single equation model, which represents current state of the practice. The mean squared predictive error and the predictive loss criteria are reduced by more than half using the mixture model, suggesting that from a statistical standpoint offers superior fit. Assessing the model for theoretical appeal, it retains all of the properties of the single equation NB model yet has the desirable property of explaining additional complexity, known to exist in crash occurrence.

The contribution of this research to the understanding of crash causation and how crashes are best understood—mainly from the theoretical questions raised—is potentially substantial. A number of particularly provocative insights are provided by this research, including:

1. For the first time the possibility of differentiating among behavioral, geometric and operational, and unobserved spatial effects through modeling is possible. An ability to predict and differentiate crashes across sites that are the result of behavioral effects versus unobserved spatial effects has enormous, if not staggering implications. Consider, for example, a procedure that would allow a safety engineer to predict that 30% of crashes at a site were the result of behavioral influences, 40% unobserved spatial influences, and the remainder a function of operational and geometric features.
2. If models were to predict counts depicted in Figure 2 instead of Figure 1, then safety investments could be strategically targeted to the appropriate kind of remediation strategies. The currently daunting task of identifying which safety investments are appropriate at various points on a network would be greatly assisted, and associated costs reduced.
3. The current hot spot identification methods, which rely on total crash counts, could be focused on where best to invest funds to address the crash components, whether operational, geometric, behavioral, or the result of spatial effects.
4. Methods that rely on safety performance functions—including those to identify hot spots—would yield entirely different results, whereby some sites may be dominated by behavioral problems while others are dominated by operational problems.

While the modeling results presented here are tremendously promising, they are not definitive, and suffer from a number of significant challenges that need to be addressed in future research. First and foremost, the mix of distributions was assumed to coincide with past research about the sources of crash causes—sources that do not directly correspondent with the components identified here. While a range of distribution mixes did not alter the results significantly, it would be important to know the correct distribution weights in any application in practice. Substantial deviations from the assumed distribution mix might yield substantially different results.

Second, and perhaps as a corollary to the first shortcoming, the distribution mix is determined endogenously in this specification and may not be sufficient to explain the between site variability as it pertains to contribution of crashes from the various components. Validation is needed to determine if exogenous mix information is sufficient to capture site to site variability in crash causes. To address this shortcoming, a study needs to be undertaken to validate, to the extent possible, whether the statistical fitting coincides with on-the-ground facts. For example, if a three component model predicts that 10 out of 30 observed crashes at a site were ‘caused’ by behavioral (random) effects, it would be important to verify that this prediction is correct. Such an exercise would require knowing with some degree of confidence the causes of crashes across a number of sites, and then estimating a model with the correct weights and validating the fit. This validation may require both experimental and observational studies.

Third, the methodology was tested on one dataset, and worked favorably. The application of the methodology to different datasets would be helpful to establish a generalizable conclusion about the underlying crash densities.

Fourth, the model here can be improved, undoubtedly. There is simultaneity of crash causes across sites that is not accommodated in this approach. Moreover, the specification of spatial effects as a function of AADT only is inferior; a better approach would be to use true spatial effects such as location and perhaps direction of travel.

Fifth, a different mix of distributions could be postulated, for example a distribution of vehicle factors could be postulated as one of the contributing distributions.

Despite these limitations, the potential impact of this different view of crash causes on our collective understanding of crash causation and practical implications on hot spot

methodologies could be profound. After validation, the implications to hot spot methods should be examined. One could imagine a procedure whereby hotspots were detected according to the type of deficiency corresponding with the underlying ‘causal mechanisms’, whereby a hotspot for behavioral deficiencies might not be the same hot spot for geometric and operational deficiencies or unobserved spatial effects.

While there is considerable further work to be done on this topic, this research has highlighted a chronic deficiency in the way crash counts are conceptualized and subsequently modeled. We observe an integration of crashes caused by fundamentally different causal mechanisms. This insight is the most important contribution of this paper. The development and validation of models that can distinctly identify and predict the contributing components of observed crash counts will lead to a breakthrough in the collective understanding of motor vehicle crashes and their remedies.

REFERENCES

1. Joshua, S. and N. Garber, *Estimating truck accident rate and involvement using linear and Poisson regression models.* Transportation Planning and Technology, 1990. **15**(1): p. 41-58.
2. Miaou, S.-P., *The relationship between truck accidents and geometric design of road sections: poisson versus negative binomial regressions.* Accident Analysis and Prevention, 1994. **26**(4): p. 471-482.
3. Vogt, A. and J.G. Bared, *Accident Models for Two-Lane Rural Roads: Segments and Intersections.* Transportation Research Record, 1998. **1635**: p. 18-29.
4. Vadlamani, S., et al., *Identifying Large Truck Hot Spots Using Crash Counts and PDOEs.* JOURNAL OF TRANSPORTATION ENGINEERING-ASCE, 2011. **137**(1): p. 11-21.
5. Poch, M. and F. Mannering, *Negative Binomial Analysis of Intersection-Accident Frequencies.* Journal of Transportation Engineering, 1996. **122**(2): p. 105-113.
6. Wang, X. and M. Abdel-Aty, *Modeling left-turn crash occurrence at signalized intersections by conflicting patterns.* Accident Analysis & Prevention, 2008. **40**(1): p. 76-88.
7. Kim, D.-G., S. Washington, and J. Oh, *Modeling crash types: new insights into the effects of covariates on crashes at rural intersections.* Journal of Transportation Engineering, 2006. **132**(4): p. 282-292.
8. Haque, M.M., H.C. Chin, and H. Huang, *Applying Bayesian hierarchical models to examine motorcycle crashes at signalized intersections.* Accident Analysis & Prevention, 2010. **42**(1): p. 203-212.
9. Graham, D., S. Glaister, and R. Anderson, *The effects of area deprivation on the incidence of child and adult pedestrian casualties in England.* Accident Analysis & Prevention, 2005. **37**(1): p. 125-135.
10. Geedipally, S.R. and D. Lord, *Investigating the effect of modeling single-vehicle and multi-vehicle crashes separately on confidence intervals of Poisson–gamma models.* Accident Analysis & Prevention, 2010. **42**(4): p. 1273-1282.
11. Lord, D. and F. Mannering, *The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives.* Transportation Research Part A: Policy and Practice, 2010. **44**(5): p. 291-305.
12. Chin, H.C. and M.A. Quddus, *Applying the random effect negative binomial model to examine traffic accident occurrence at signalized intersections.* Accident Analysis and Prevention, 2003. **35**(2): p. 253-259.
13. Wang, X. and M. Abdel-Aty, *Temporal and spatial analyses of rear-end crashes at signalized intersections.* Accident Analysis & Prevention, 2006. **38**(6): p. 1137-1150.
14. Mitra, S. and S. Washington, *On the nature of over-dispersion in motor vehicle crash prediction models.* Accident Analysis and Prevention, 2007. **39**(3): p. 459-468.
15. Anastasopoulos, P.C. and F.L. Mannering, *A note on modeling vehicle accident frequencies with random-parameters count models.* Accident Analysis & Prevention, 2009. **41**(1): p. 153-159.
16. Shankar, V., J. Milton, and F. Mannering, *Modeling accident frequencies as zero-altered probability processes: an empirical inquiry.* Accident Analysis and Prevention, 1997. **29**(6): p. 829-837.

17. Kumara, S.S.P. and H.C. Chin, *Modeling accident occurrence at signalized tee intersections with special emphasis on excess zeros.* Traffic Injury Prevention, 2003. **4**(1): p. 53-57.
18. Lord, D., S. Washington, and J.N. Ivan, *Further notes on the application of zero-inflated models in highway safety.* Accident Analysis & Prevention, 2007. **39**(1): p. 53-57.
19. Lord, D., S.P. Washington, and J.N. Ivan, *Poisson, Poisson-gamma and zero-inflated regression models of motor vehicle crashes: balancing statistical fit and theory.* Accident Analysis & Prevention, 2005. **37**(1): p. 35-46.
20. Malyshkina, N.V., F.L. Mannering, and A.P. Tarko, *Markov switching negative binomial models: An application to vehicle accident frequencies.* Accident Analysis & Prevention, 2009. **41**(2): p. 217-226.
21. Park, B.-J. and D. Lord, *Application of finite mixture models for vehicle crash data analysis.* Accident Analysis & Prevention, 2009. **41**(4): p. 683-691.
22. Ma, J., K.M. Kockelman, and P. Damien, *A multivariate Poisson-lognormal regression model for prediction of crash counts by severity, using Bayesian methods.* Accident Analysis & Prevention, 2008. **40**(3): p. 964-975.
23. Park, E.S. and D. Lord, *Multivariate Poisson-lognormal models for jointly modeling crash frequency by severity.* Transportation Research Record, 2007. **2019**: p. 1-6.
24. Ye, X., et al., *A simultaneous equations model of crash frequency by collision type for rural intersections.* Safety Science, 2009. **47**(3): p. 443-452.
25. Rumar, K., ed. *The role of perceptual and cognitive filters in observed behavior*. Human Behavior in Traffic Safety, eds. L. Evans and R. Schwing. 1985, Plenum Press.
26. Aguero-Valverde, J. and P.P. Jovanis, *Spatial analysis of fatal and injury crashes in Pennsylvania.* Accident Analysis & Prevention, 2006. **38**(3): p. 618-625.
27. Mitra, S. and S. Washington, *On the significance of omitted variables in intersection crash modeling.* Accident Analysis & Prevention, 2012. **In press**.
28. Washington, S.P., M.G. Karlaftis, and F.L. mannering, *Statistical and Econometric Methods for Transportation Data Analysis*. 2003, Boca Raton, FL: Chapman and Hall/CRC.
29. Oh, J., et al., *Validation of the FHWA crash models for rural intersections: lessons learned.* Transportation Research Record, 2003. **1840**: p. 41-49.
30. Sabey, B. and G. Staughton, *Interacting roles of road environment, vehicle, and road user*, in *In Proc. 5th International Association for Accident and Traffic Medicine.* 1975: London.
31. Gelfand, A. and S. Ghosh, *Model Choice: A Minimum Posterior Predictive Loss Approach.* Biometrika, 1998. **85**: p. 1-11.
32. Abdel-Aty, M.A. and A.E. Radwan, *Modeling traffic accident occurrence and involvement.* Accident Analysis and Prevention, 2000. **32**(5): p. 633-642.
33. Meliker, J.R., et al., *Spatial analysis of alcohol-related motor vehicle crash injuries in southeastern Michigan.* Accident Analysis & Prevention, 2004. **36**(6): p. 1129-1135.